

Problem 1

Part A

Note that $|Y_i| \leq \delta n$, thus $Y_i^2 \leq \delta^2 n^2$. Therefore

$$\text{Var}(Y_i) = \mathbb{E}[Y_i^2] - (\mathbb{E}[Y_i])^2 \leq \mathbb{E}[Y_i^2] \leq \delta^2 n^2$$

which was to be shown.

Part B

Let $Y = \frac{1}{n}(Y_1 + \dots + Y_n)$. Note that

$$\text{Var}(Y) = \frac{1}{n^2} \text{Var}(Y_1 + \dots + Y_n) \leq \frac{1}{n^2} \cdot n \cdot \delta^2 n^2 = \delta^2 n.$$

Therefore by Chebychev's,

$$\mathbb{P}[|Y - \mathbb{E}[Y]| > \varepsilon] \leq \frac{\delta^2 n}{\varepsilon} = \frac{n}{n^{\frac{4}{3}} \varepsilon} = \frac{1}{n^{\frac{1}{3}} \varepsilon}.$$

The upper bound goes to 0 as $n \rightarrow \infty$, thus in the limit the probability is equal to 0.

Part C

Note that

$$\delta n \cdot \mathbb{P}[|X| > \delta n] \leq \delta n \cdot \sum_{|x| > \delta n} \mathbb{P}[|X| = x] \leq \sum_{|x| > \delta n} x \mathbb{P}[X = x] \leq \sum_{x > \delta n} x \mathbb{P}[X = x].$$

Thus

$$\mathbb{P}[|X| > \delta n] \leq \frac{1}{\delta n} \cdot \sum_{x > \delta n} x \mathbb{P}[X = x].$$

Part D

Let

$$L_n = \sum_{x \leq n} x \mathbb{P}[X = x] \quad R_n = \sum_{x > n} x \mathbb{P}[X = x].$$

That is, L_n is the left tail of the expectation of X and R_n is the right tail. Note that since the expectation is finite, $\lim_{n \rightarrow \infty} L_n = \mathbb{E}[X]$ and $L_n + R_n = \mathbb{E}[X]$. Therefore as the right tail cutoff approaches infinity, in the limit it must go to 0. Choosing $\delta = \frac{1}{n^{2/3}}$ gives then

$$\lim_{n \rightarrow \infty} \sum_{x > \delta n} x \mathbb{P}[X = x] = \lim_{n \rightarrow \infty} \sum_{x > n^{1/3}} x \mathbb{P}[X = x] = 0.$$

Part E

By the union bound

$$\mathbb{P}[\exists Z_i \text{ s.t. } Z_i \neq 0] \leq n \cdot \mathbb{P}[|X| > \delta n] \leq \frac{1}{\delta} \sum_{x > \delta n} x \mathbb{P}[|X| = x].$$

From (d), it follows this upper bound goes to 0 as $n \rightarrow \infty$.

Part F

Note that $X = Y_i + Z_i$. Therefore

$$\mathbb{E}[X] = \mathbb{E}[Y_i] + \mathbb{E}[Z_i].$$

Since $Z_i = 0$ with high probability from (e), it follows that

$$\lim_{n \rightarrow \infty} \mathbb{E}[Y_i] + \mathbb{E}[Z_i] = \lim_{n \rightarrow \infty} \mathbb{E}[Y_i] = \mathbb{E}[X].$$

Part G

Note that

$$\mathbb{E}[Y] = \frac{1}{n} \cdot n \cdot \mathbb{E}[Y_1] = \mathbb{E}[Y_1] = \mathbb{E}[X].$$

In the limit $n \rightarrow \infty$ it then follows from (f) and (b)

$$\lim_{n \rightarrow \infty} \mathbb{P}\left[\left|\frac{\sum_{i=1}^n X_i}{n} - \mathbb{E}[X]\right| > \varepsilon\right] = \lim_{n \rightarrow \infty} \mathbb{P}\left[\left|\frac{\sum_{i=1}^n Y_i}{n} - \mathbb{E}[Y]\right| > \varepsilon\right] = 0.$$

Problem 2

Proof. Note that the moment generating functions of X and Y are

$$M_X(t) = \exp\left(\mu_X t + \frac{1}{2}\sigma_X^2 t^2\right) \quad M_Y(t) = \exp\left(\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2\right).$$

Since X and Y are independent, the moment generating function of their sum is the product

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \exp\left((\mu_X + \mu_Y)t + \frac{1}{2}(\sigma_X^2 + \sigma_Y^2)t^2\right).$$

Note that this is the same moment generating function of a normal distribution with mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$. Thus by uniqueness of moment generating functions, this is the distribution of $X + Y$. \diamond

Problem 3

Let I_{ij} be the indicator random variable that a_i and a_j are inverted. The probability that a_i and a_j are inverted is the probability that $\pi(i) > \pi(j)$ where π is the random permutation. This is simply $\frac{1}{2}$, thus $\mathbb{P}[I_{ij} = 1] = \frac{1}{2}$. Note that a given iteration of bubble sort removes exactly 1 inversion, thus the number of inversions that bubblesort will correct is the number of inversions in the original permutation, which is equal to

$$\mathbb{E}\left[\sum_{i=1}^n \sum_{j>i}^n I_{ij}\right] = \sum_{i=1}^n \sum_{j>i}^n \mathbb{E}[I_{ij}] = \sum_{i=1}^n \sum_{j>i}^n \frac{1}{2} = \binom{n}{2} \cdot \frac{1}{2} = \frac{n(n-1)}{4}.$$

Note that the first pass moves the largest element to the top, second largest before the top, etc. This means that on the first pass, $n-1$ comparisons are done, on the second pass one less comparison is done meaning $n-2$ comparisons, etc. Therefore the expected number of comparisons is

$$(n-1) + (n-2) + \dots + 1 = \frac{n(n-1)}{2}.$$

Problem 4

Let X_{ij} be the indicator random variable that the i th smallest and j th smallest elements are compared. Since quicksort sorts sublists, if two elements end up in two separate sublists, they will never be compared. Therefore the total number of comparisons X can be expressed as

$$X = \sum_{i=1}^n \sum_{j>i}^n X_{ij}.$$

Note that the only places for the pivot that affects X_{ij} is between i and j , since if the pivot is out of this range then both are in the same sublist and so a pivot will be reselected again. X_{ij} is only 1 when the pivot is selected to be either the element associated with i or j , as otherwise the two elements would be in different sublists and hence never compared. Thus $\mathbb{P}[X_{ij} = 1] = \frac{2}{j-i+1}$. Therefore

$$\mathbb{E}[X] = \sum_{i=1}^n \sum_{j=i+1}^n \mathbb{E}[X_{ij}] = 2 \sum_{i=1}^n \sum_{j=2}^{n-i+1} \frac{1}{j} < 2 \sum_{i=1}^n \sum_{j=2}^n \frac{1}{j} = 2n(H_n - 1)$$

where H_n is the n th harmonic number. Note that $\log n \leq H_n \leq \log n + 1$ thus $\mathbb{E}[X] \leq 2n \log n$. Therefore by Markov's inequality,

$$\mathbb{P}[X \geq cn \log n] \leq \frac{\mathbb{E}[X]}{cn \log n} \leq \frac{2n \log n}{cn \log n} = \frac{2}{c}.$$

Choosing $c = 200$

$$\mathbb{P}[X < cn \log n] = 1 - \mathbb{P}[X \geq cn \log n] \geq 1 - \frac{2}{c} = 1 - \frac{1}{100} = \frac{99}{100}$$

which was to be shown.

Problem 5

- a) This random variable most likely would not be normally distributed since the score distribution for each question is not guaranteed to be

similar. However if the exam was designed such that the score distribution for each question is roughly the same, they are independent and so central limit theorem would apply, giving approximate normality

- b) This random variable should be normally distributed as it is a sum of many i.i.d Bernoulli random variables, thus by the central limit theorem the average of these variables is normal, hence the sum of them is also normally distributed
- c) This random variable should not be normally distributed, as height of men is likely distributed differently than women meaning each sample is not identically distributed meaning (its going to be more bimodal). Therefore the central limit theorem doesn't guarantee normality
- d) This random variable should be normally distributed as it doesn't have the problem meeting the requirements for the central limit theorem that were discussed in (c). The height of a woman is i.i.d to any other woman, meaning the average of all these heights should be approximately normal